A Unifying Model for the Analysis of Phenotypic, Genetic and Geographic Data

Guillot, Gilles; Rena, Sabrina; Ledevin, Ronan; Michaux, Johan; Claude, Julien

Published in:
Systematic Biology

Link to article, DOI:
10.1093/sysbio/sys038

Publication date:
2012

Document Version
Early version, also known as pre-print

A Unifying Model for the Analysis of Phenotypic, Genetic and Geographic Data

Gilles Guillot*†, Sabrina Renaud‡, Ronan Ledevin‡, Johan Michaux§, Julien Claude¶

December 21, 2011

*Corresponding author; gigu [at] imm.dtu.dk.

†Informatics Department, Technical University of Denmark, Copenhagen, Denmark.

‡Laboratoire de Biométrie et Biologie Evolutive UMR 5558, CNRS, Université Lyon 1, Université de Lyon, 69622 Villeurbanne, France,

§CBGP, INRA/IRD/CIRAD/SupAgro, Montferrier-sur-Lez cedex, France.

¶Laboratoire de Morphométrie, ISE-M, UMR5554 CNRS/UM2/IRD, Université de Montpellier II, 34095 France.

Abstract

Recognition of evolutionary units (species, populations) requires integrating several kinds of data such as genetic or phenotypic markers or spatial information, in order to get a comprehensive view concerning the differentiation of the units. We propose a statistical model with a double original advantage: (i) it incorporates information about the spatial distribution of the samples, with the aim to increase inference power and to relate more explicitly observed patterns to geography; and (ii) it allows one to analyze genetic and phenotypic data within a unified model and inference framework, thus opening the way to robust comparisons between markers and possibly combined analyzes. We show from simulated data as well are real data from the literature that our method estimates parameters accurately and improves alternative approaches in many situations. The interest of this method is exemplified using an intricate case of inter- and intra-species differentiation based on an original data-set of georeferenced genetic and morphometric markers obtained on *Myodes* voles from Sweden. A computer program is made available as an extension of the R package Geneland.

Keywords

Clustering, spatial data, bio-geography, Bayesian model, Markov chain Monte Carlo, R package, morphometrics, molecular markers, *Myodes*.

Species delimitation are of interest in conservation biology (identification and management of endangered species), epidemiology (detection of new pathogens) but also from a purely cognitive point of view to describe, quantify and understand mechanisms of speciation. Methodological advances in evolutionary biology have led to methods for species identification solely based on the variation of key genetic markers (e.g. DNA barcoding, Luo et al., 2011). Limits of these single-marker approaches are more and more evidenced by conflicts between different genes in a multi-marker approach (Rodríguez et al., 2010; Turmelle et al., 2011) or between genetic and phenotypic markers (Nesi et al., 2011). In this context of species or population identification, phenotypic data still emerge of interest together with genetic markers.

Phenotypic data such as size and/or shape of morphological structures are the product of numerous interacting nuclear genes (Klingenberg et al., 2001) and as such can provide a global estimate of the divergence between units. Furthermore, by being the target of the screening by selection, morphological variation can provide precious insights on the selection pattern contributing to shape the units. In the case of fossil lineages, it may even be the only information available to identify evolutionary and systematic units (Néraudeau, 2011; Girard and Renaud, 2011).

A rich toolbox is available to tackle these questions. Many methods work as partition clustering, and aim at defining how many groups are represented in a sample of individuals, and assign these individuals to these groups following some optimality principles. These methods were initially developed to deal with continuous quantitative measurements. These classical clustering methods have been implemented in programs such as Emmix (McLachlan et al., 1999) or Mclust (Fraley, 1999) or Mixmod (Biernacki et al., 2006). The methods above did not received a strong interest in Systematics until recent Population Genetics extensions to deal with molecular data such as the widely used computer program Structure (Pritchard et al., 2000) and related work (reviewed e.g. by Excoffier and Heckel, 2006). More recently, Hausdorf and Hennig (2010) and

Yang and Rannala (2010) developed methods for delimiting species based on multi-locus data. While the approach of Hausdorf and Hennig (2010) method hinges on Gaussian clustering, the method of Yang and Rannala (2010) is based on the coalescent and makes use of a user-specified guide tree. Methods for genetic data have been also extended to incorporate information about the spatial location of each sample - an information rarely used although commonly available in data analysis in evolutionary biology - with the aim of increasing power of inferences and of relating more explicitly observed patterns to geography (Guillot et al., 2005, 2009).

These tools have been developed by different communities (evolutionists, population geneticists, statisticians). Therefore, one still lacks a unified framework, and this constitutes a major drawback for combining various kinds of data. This is especially true for morphological markers that did not received as much attention as genetic markers for recognizing populations and species. There are therefore a few major gaps in the toolbox available to identify evolutionary units, namely there is to date: no method to analyze genetic data and phenotypic data under the same general paradigm (model and inference framework), and no method to incorporate spatial information in such phenotypic/genetic analysis.

The goal of the present paper is to fill these gaps. We propose a model to deal in an integrated way with georeferenced phenotypic and genetic data and we provide a computer program freely available that implements this model and should ease data analysis in many respects. Given the complexity of the modeling and inferential task, our method is not based on an explicit evolutionary model (for example based on the coalescent) but on a statistical model. This model is a parametrization which is general enough to capture some essential features in the data variation, but also simple enough to be subject to a rigorous and accurate inference method. Briefly, our model assumes the existence of several clusters which display some kind of homogeneity. This model mimics more or less what would be expected from a population: homogeneity in

terms of genetic and phenotypic variation and some geographical continuity. The existence of homogeneous clusters corresponds to the fact that some individuals have shared some aspects of their recent ecological or evolutionary history. This shared history is summarized by cluster-specific parameters which are allele frequencies and means and variances of phenotypic traits. Because it is not based on an explicit evolutionary model, it does not require prior information (as for instance a guide tree in the case of Yang and Rannala's method). The statistical challenge in this context is to estimate the number of clusters and these cluster-specific parameters.

This article is organized as follows. First we provide a description of the model and inference machinery. Next we illustrate our method and test its accuracy on a large set of simulated data as well as on two published real data-sets. Then we implement our method on an original data-set of georeferenced genetic and morphometric markers to decipher the complex inter-and intra-specific structure of red-backed and bank voles *Myodes rutilus* and *M. glareolus* in Sweden. We conclude by discussing potential applications in a more general context.

Method

Overview

We assume that we have a data-set consisting of n individuals sampled at sites $\mathbf{s} = (s_i)_{i=1,...,n}$ (where s_i is the two-dimensional spatial coordinate of individual i), observed at some phenotypic variables denoted $\mathbf{y} = (y_{ij})_{\substack{i=1,...,n \\ j=1,...,q}}$ and/or some genetic markers denoted $\mathbf{z} = (z_{ij})_{\substack{i=1,...,n \\ j=1,...,l}}$. Our approach is able to deal with any combination of phenotypic and genetic data, including situations where only phenotypic or only genetic data are available and situations when each individual is observed through its own combination of phenotypic and genetic markers. As it will be shown below, our approach also encompasses the case where sampling locations are missing (or considered to be irrelevant). The only constraint that we impose at this stage is that if spatial coordinates are used, they must be available for all individuals. We assume that each individual sampled belongs to one of K different clusters and that variation in the data can be captured by cluster-specific location and scale parameters.

Prior and Likelihood Model for Phenotypic Variables

Denoting by p_i the cluster membership of individual i ($p_i \in \{1, ..., K\}$), we assume that conditionally on $p_i = k$, y_{ij} is drawn from a parametric distribution with cluster-specific parameters. Independence is assumed within and across clusters conditionally on cluster membership. This means in particular that there is no residual dependence between variables not captured by cluster memberships. Implications of this assumption are discussed later. Although most of the analysis that follows would be valid for all families of continuous distribution, we assume in the following that the y values arise from a normal distribution. Each cluster is therefore characterized by a mean μ_{kj} and a variance σ^2_{kj} and our model is a mixture of multivariate independent normal distributions (Frühwirth-Schnatter, 2006). Following a common practice in Bayesian analysis

(Gelman et al., 2004), we use the natural conjugate prior family on $(\mu_{kj}, 1/\sigma^2_{kj})$ for each cluster k and variable j. Namely, we assume that the precision $1/\sigma^2_{kj}$ (i.e. inverse variance) follows a Gamma distribution $\mathcal{G}(\alpha, \beta)$ (α shape, β rate parameter) and that conditionally on σ_{kj}, the mean μ_{kj} has a normal distribution with mean ξ and variance σ^2_{kj}/κ. In the specification above, α, β, ξ and κ are hyper-parameters. Details about their choice are discussed in the appendix and in the supplementary material.

Prior and Likelihood Model for Genetic Data

We assume here a mixture of multinomial distributions. This is the model previously introduced by Pritchard et al. (2000) to model individuals with pure ancestries. Denoting frequency of allele a at locus l in cluster k by f_{kla}, for diploid genotype data we assume that

$$\pi(z_{ij} = \{a, b\} | p_i = k) = 2 f_{kla} f_{klb} \quad \text{whenever} \quad a \neq b \tag{1}$$

$$\text{and } \pi(z_{ij} = \{a, a\} | p_i = k) = f^2_{kla}. \tag{2}$$

While for haploid data, we have

$$\pi(z_{ij} = a | p_i = k) = f_{kla} \tag{3}$$

We also deal with dominant markers for diploid organisms with a modified likelihood (see Guillot and Santos, 2010; Guillot and Carpentier-Skandalis, 2011, for details). We assume independence of the various loci within and across clusters conditionally on cluster memberships. In particular, as with all other population genetic clustering models (including STRUCTURE), we do not attempt to model background linkage disequilibrium (LD). Therefore, our model can handle non-recombining DNA sequences (such as data obtained from mitochondrial DNA, Y chromosomes or tightly linked autosomal nuclear markers) provided data are reformatted in such a way that the various haplotypes are recoded as alleles of a single locus, but see also discussion. We assume that

allele frequencies $f_{kl.}$ have a Dirichlet distribution. Independence of the vectors $f_{kl.}$ is assumed across loci. Regarding the dependence structure across clusters, we consider either independence (referred to as Uncorrelated Frequency Model or UFM) or an alternative model (referred to as Correlated Frequency Model or CFM) introduced by Balding and Nichols (1995, 1997). In this second model, allele frequencies also follow a Dirichlet distribution but now depending on some cluster-specific drift parameters. In this model, $f_{kl.}$ are assumed to follow a Dirichlet distribution $\mathcal{D}(\tilde{f}_{la}(1-d_k)/d_k, ..., \tilde{f}_{lA}(1-d_k)/d_k)$ where d_ks parametrize the speed of divergence of the various clusters and the $\tilde{f}_{la}$s represent the allele frequency in an hypothetical ancestral population. This model can be viewed as a heuristic and computationally convenient approximation of a scenario in which present time clusters result from the split of an ancestral cluster some generations ago. It is also a Bayesian way of introducing correlation between clusters at the allele frequency level and hence to infer subtle differentiations that would have been missed by a model assuming independence of allele frequencies across clusters (Falush et al., 2003; Guillot, 2008; Sirén et al., 2011) .

Prior Models for Cluster Membership

Spatial model

We consider a statistical model known as colored Poisson-Voronoi tessellation. Loosely speaking, this model assumes that each cluster area in the geographic domain can be approximated by the union of a few polygons. Most of the modeling ideas can be grasped from the examples shown in figure 1. The polygons are assumed to be centered around some points that are generated by a homogeneous Poisson process (i.e. points located completely at random in the geographic domain). Formally, we denote by $(u_1, ..., u_m)$ the realization of this Poisson process. These points in $\mathbb{R}^2$ induce a Voronoi tessellation into m subsets $\Delta_1, ..., \Delta_m$. The Voronoi tile associated with point u_i is defined as $\Delta_i = \{s \in \mathbb{R}^2, dist(s, u_i) < \text{dist}(s, u_j) \forall j \neq i\}$. Each tile receives a

cluster membership c_i (coded graphically as a color hence the terminology) at random sampled independently from a uniform distribution on $\{1, ..., K\}$. Denoting by D_k the union of tiles with color k, the set $(D_1, ..., D_K)$ defines a tessellation in K subsets. This model is controlled by the intensity of the Poisson process λ (the average number of points per unit area) and the number of clusters K. We place a uniform prior on $[0, \lambda_{max}]$ and on $\{0, ..., K_{max}\}$ respectively. This model is a flexible tool widely used in engineering to fit arbitrary shapes in a non-parametric way (Møller and Stoyan, 2009). It offers a good trade-off between model complexity, realism and computational efficiency. It is presumably most useful in situations of incipient allopatric speciation but examples of applications in other contexts can be found e.g. in the studies of Coulon et al. (2006); Fontaine et al. (2007); Wasser et al. (2007); Hannelius et al. (2008); Joseph et al. (2008); Sacks et al. (2008); Galarza et al. (2009); Beadell et al. (2010). See also Guillot et al. (2009) for review and additional references. Lastly, we note that our approach relates to that of Hausdorf and Hennig (2003) who propose a test for clustering of areas of distribution. However, rather than testing clusteredness, our approach estimates these areas of distribution. To do that, we assume some clusteredness but without making strong assumptions about its intensity.

Non-spatial model

If spatial coordinates are not available or thought to be irrelevant to the species at the spatial scale considered, then a non-spatial model can be used. The non-spatial modeling option considered here does not require to introduce any auxiliary point process as above but for the sake of consistency, we use the same setting as in the paragraph above. We set $m = n$ and impose $(u_1, ..., u_n) = (s_1, ..., s_n)$. Here the s_i s are some known spatial coordinates or dummy points if this piece of information is missing. This model does not impose any spatial structure and corresponds to the model implemented in most non-spatial cluster programs, including the genetic clustering programs Baps (Corander et al., 2003, 2004) and Structure (with the exception of the latest

model presented by Hubisz et al. (2009).

[Figure 1 about here.]

Summary of Proposed Model

The parameters in our model are as follows: number of clusters K, rate of Poisson process λ, number of events (points) of the Poisson process m, events of Poisson process $\mathbf{u} = (u_1, ..., u_m)$, color of tiles (i.e. cluster membership of spatial partitioning sub-domains) $\mathbf{c} = (c_1, ..., c_m)$, allele frequencies $\mathbf{f} = (f_{kla})$ (frequency of allele a at locus l in cluster k), genetic drift parameters $\mathbf{d} = (d_1, ..., d_K)$, allele frequencies in the ancestral population $\tilde{\mathbf{f}} = (\tilde{f}_{la})$, expectations of phenotypic variables $\boldsymbol{\mu} = (\mu_{kj})$, standard deviations of phenotypic variables $\boldsymbol{\sigma} = (\sigma_{kj})$ (note that $\boldsymbol{\sigma}$ is not a variance-covariance matrix (the phenotypic variables are assumed to be independent) but rather a set of scalar variances stored in a two-dimensional array. On top of this, we place a uniform prior on $[0, \lambda_{max}]$ on λ, a uniform prior on $\{0, ..., K_{max}\}$ on K, a Beta $\mathrm{B}(\delta_k, \delta_k)$) prior on d_k and a Gamma distribution $\mathcal{G}(g, h)$ on β.

The vector of unknown parameters is therefore $\boldsymbol{\theta} = (K, \lambda, m, \mathbf{u}, \mathbf{c}, \mathbf{f}, \tilde{\mathbf{f}}, \mathbf{d}, \boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\beta})$. We also denote by $\boldsymbol{\theta}_S = (\lambda, m, \mathbf{u}, \mathbf{c})$, $\boldsymbol{\theta}_G = (\mathbf{f}, \tilde{\mathbf{f}}, \mathbf{d})$ and $\boldsymbol{\theta}_P = (\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\beta})$ the parameters of the spatial, genetic and phenotypic parts of the model respectively.

The hierarchical structure of the model is summarized on the graph shown in figure 2. There are three blocks of parameters relative to the genetic, phenotypic and geographic component of the model. Information propagates from data to higher levels of the model across the various nodes of the graph through probabilistic relationships specified between neighboring nodes.

[Figure 2 about here.]

The structure of the global model can be summarized by the joint distribution of $\boldsymbol{\theta}$ and $(\mathbf{y}, \mathbf{z})$.

By the conditional independence assumptions, we get

$$\begin{aligned} \pi(\boldsymbol{\theta}, \mathbf{y}, \mathbf{z}) &= \pi(\boldsymbol{\theta})\pi(\mathbf{y}, \mathbf{z}|\boldsymbol{\theta}) \\ &= \pi(\boldsymbol{\theta})\pi(\mathbf{y}|\boldsymbol{\theta})\pi(\mathbf{z}|\boldsymbol{\theta}) \\ &= \pi(\boldsymbol{\theta})\pi(\mathbf{y}|\boldsymbol{\theta}_P)\pi(\mathbf{z}|\boldsymbol{\theta}_G) \end{aligned} \tag{4}$$

Each genetic or phenotypic marker brings one factor in the likelihood. Whether the clustering is driven by the genetic or the phenotypic data depends on the respective differentiation and on the number of markers of each kind.

Estimation of Parameters

Bayesian estimation and Markov chain Monte Carlo inference

We are interested in the posterior distribution $\pi(\boldsymbol{\theta}|\mathbf{y}, \mathbf{z})$. Note that this notation does not refer explicitly to the sample locations because, unlike genetic markers and phenotypic variables, locations are not considered as random quantities in our model. The model does in fact implicitly account for spatial information. The distribution $\pi(\boldsymbol{\theta}|\mathbf{y}, \mathbf{z})$ is defined on a high dimensional space and deriving properties analytically about this distribution is out of reach. We implement a Markov chain Monte Carlo strategy. This amounts to generating a sample of N correlated replicates $(\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_N)$ from the posterior distribution $\pi(\boldsymbol{\theta}|\mathbf{y}, \mathbf{z})$. The initial state $\boldsymbol{\theta}_1$ is simulated at random from a distribution that does not matter in principle, a fact that has to be checked in practice by convergence monitoring tools (Gilks et al., 1996; Robert and Casella, 2004). We always sample $\boldsymbol{\theta}_1$ from the prior and we check that starting from various random states does not affect the overall result provided a suitable number of burn-in iterations are discarded. In analyzes reported below, the order of magnitude of N was 50000-100000 iterations with 20000 burn-in iterations. See appendix for detail on the MCMC algorithm.

Estimation of the number of clusters

Each simulated state $\boldsymbol{\theta}_i$ includes a simulated number of clusters K_i. The number of clusters is estimated as the most frequent value among the N simulated values $K_1, ..., K_N$ and we denote it by $\hat{K}$.

Estimating cluster memberships

A model assuming that individuals i and j belong respectively to clusters 1 and 2 characterized by a mean phenotypic trait equal to 5 and 7 is essentially the same as a model assuming that individuals i and j belong respectively to clusters 2 and 1 characterized by a mean phenotypic trait equal to 7 and 5. This trivial fact is due to the invariance of the likelihood under permutation of cluster labels and brings up a number of computational difficulties in the post-processing of MCMC algorithm outputs known as the label switching issue (Stephens, 1997). In particular, it does not make sense to average values across the MCMC iterations. To deal with this, we implement the strategy described by Marin et al. (2005) and Guillot (2008). We consider the set of simulated $\boldsymbol{\theta}$ values restricted to the set of states such that $K = \hat{K}$. Then working on this restricted set, we relabel each state in such a way that they "best look like" the modal state of the posterior distribution. Cluster memberships of each individual are estimated as the modal value in this relabeled sample. Then we estimate all cluster-specific parameters (mean phenotypic values and allele frequencies) by taking the average simulated value over the relabeled sample.

Analysis of Simulated Data

We investigate here two new aspects of the model, namely its ability to cluster phenotypic data only and phenotypic and genetic data jointly together with some spatial information.

Inference from Phenotypic Data Only

In this section, we present new results on the model for phenotypic data and focus on the spatial model option. We carried out simulations from our prior model and performed inferences as described in section "Estimation of parameters" above. We produced data-sets consisting of $n = 200$ individuals with $q = 5, 10, 20$ and 50 phenotypic variables. For each value of q, we produced 500 data-sets with a uniform prior $\mathcal{U}(\{1, ..., 5\})$ on K. In real-life, the range of value of the putative true K is largely unknown. To be as close as possible to this situation, we carried out inference under a uniform $\mathcal{U}(\{1, ..., 10\})$ prior for K. We assessed the accuracy of inferences by computing the classification error which is displayed in figure 3. Further details are provided in Supporting Material.

We also wished to assess how our method performs compared to other computer programs implementing state-of-the-art methods. We therefore considered the R package MCLUST (Banfield and Raftery, 1993; Fraley, 1999) which is one of the most widely used and arguably most advanced program to perform clustering. This program implements inference for Gaussian mixtures and as such deals solely with continuous quantitative data. It implements a non-spatial algorithm and in its default setting performs inference by likelihood maximization via the Expectation Maximization (EM) algorithm. It implements a wide class of sub-models regarding the covariance structure of the data. In its default option (which we used) it performs model selection (covariance structure and number of clusters) by optimizing a Bayesian Information Criterion (BIC). We set the maximum number clusters to the $K_{max} = 10$, i.e. to the same value as in analyzes with our method.

We stress here that the goal of this experiment is not to rank our method and MCLUST as the two methods/programs differ in many important respects. They differ regarding the type of data handled (MCLUST is not aimed at genetic data and does not implement any spatial model) and

the breadth of covariance structure considered (our approach assumes conditional independence while MCLUST considers in excess of ten types of covariance structures). It would be therefore difficult to design an efficient and fair comparison. Results are mostly given here to support the claim that our method compares with state-of-the-art methods and to assess the magnitude of improvement brought by the use of a spatial model in a best-case scenario when data are spatially structured (see also discussion). Most of the numerical results are summarized in figure 3.

[Figure 3 about here.]

To understand better how the method behaves as a function of the pairwise phenotypic differentiation between clusters, we also report the classification error as a function of the T^2 statistic in a Hotelling T test (Anderson, 1984) on figure 4. See also supporting material for further details.

[Figure 4 about here.]

Inference from Phenotypic and Genetic Data jointly

We illustrate here how combining phenotypic and genetic data can improve the accuracy of inferences compared to inferences carried out from one type of data only. To do so, we simulated 500 data-sets consisting of two clusters each. There were five phenotypic variables and ten co-dominant genetic markers. We investigated a broad range of phenotypic and genetic differentiation and it appears that on average combining the two types of data increases the accuracy of inferences. See figure 5.

[Figure 5 about here.]

Analysis of Data from the Literature

Analysis of Iris Morphometric Data

Fisher's iris data-set (Anderson, 1935; Fisher, 1936) gives the measurements in centimeters of the variables sepal length and width and petal length and width, respectively, for 50 flowers from each of 3 species of iris. The species are *Iris setosa*, *versicolor*, and *virginica*. We applied our method to the data transformed into log shape ratios (see Claude, 2008, and references therein). Since the data are not georeferenced, we used the non-spatial prior. We launched ten independent MCMC runs. Seven of them return correctly $\hat{K} = 3$, the other three runs return $\hat{K} = 4, 5$ and 6 respectively. Ranking the runs according to the average posterior density, the best run corresponds to one of the seven runs that estimate K correctly (according to the number of actual species in the data set). This run achieves a classification error of 6% (see Fig. 6). Mclust returns an estimate of K equal to 2 (raw data or log shape ratio data) and 50 out of 150 individuals are misclassified, thus failing to identify the three species of the data set.

[Figure 6 about here.]

AFLP Data of *Calopogon* from Eastern North America and the Northern Caribbean

The way our model deals with genetic data and the accuracy resulting from this method based on genetic data only has been investigated by Guillot et al. (2005, 2008); Guillot (2008); Guillot and Santos (2010); Safner et al. (2011); Guillot and Carpentier-Skandalis (2011) and further discussion can be found in Guillot et al. (2009), however, to further illustrate the accuracy of our method when used with genetic data only, we study here a dataset produced and first analyzed by Goldman et al. (2004).

This dataset consists of sixty *Calopogon* samples genotyped at 468 AFLP markers. Goldman et al. (2004) identified the presence of five species (*C. barbatus, C. oklahomensis, C. tuberosus, C.*

pallidus, C. multiflorus) and two hybrids specimens (*C. tuberosus* × *C. pallidus* and *C. pallidus* × *C. multiflorus*). According to Goldman et al. (2004), *C. tuberosus* has been widely considered to have three varieties: var. *tuberosus*, var. *latifolius* and var. *simpsonii*. In addition, the dataset contains samples from two outgroups so that one could consider that the dataset contains up to eleven distinct species.

We analysed this dataset under the same setting as the previous dataset. Under the UFM, the estimated K ranges between 2 and 3 . The best run (in terms of average posterior density) corresponds to $\hat{K} = 3$. In this clustering, one cluster contains the samples of the *C. tuberosus* species, a second cluster merges the samples of the *C. barbatus, C. oklahomensis, C. pallidus, C. multiflorus* species and the hybrids. The last cluster contains the samples from the two outgroups. Under the CFM, the estimated K ranges between 7 and 8 . The best run (in terms of average posterior density) corresponds to $\hat{K} = 8$. It clusters the individuals of the various species as follows: *C. oklahomensis* / *C. multiflorus* / *C. barbatus* / *C. pallidus*, *C. tuberosus* × *C. pallidus* and *C. pallidus* × *C. multiflorus* / *C. tuberosus tuberosus* except three samples / the three *C. tuberosus tuberosus* previous samples / two extra clusters for the outgroups.

Analysis of *Myodes* Vole Data

Data and statistical analysis

We now study an original dataset of geo-referenced genetic and phenotypic markers of the voles of the genus *Myodes* in Sweden. This dataset has several interests to investigate the efficiency of our method on a complex real case. (i) Fennoscandia has been recognised as a zone where the mitochondrial DNA of the northern red-backed vole *Myodes rutilus* introgressed its southern relative, the bank vole *M. glareolus* (Tegelström, 1987). This makes the identification of these two species impossible based on common mitochondrial markers. (ii) The bank vole is further

characterized by intra-specific lineages (Deffontaine et al., 2009). Two of them are documented in Sweden (Razzauti et al., 2009), providing a complex case for disentangling intra- and inter-specific structure. (iii) Both genetic and morphological data are available on this model to confront the structure provided by the two kinds of markers, and test for their combination.

The dataset consists of 182 individuals. These individuals were genotyped at 14 microsatellite loci (Lehanse, 2010). The phenotypic dataset corresponds to a subsample of 69 individuals (Ledevin, 2010). We used measurements of the third upper molar shape, for which a phenotypic differentiation has been evidenced at the phylogeographic scale (Deffontaine et al., 2009; Ledevin et al., 2010a). The two-dimensional outline was manually registered from numerical pictures, starting from a comparable starting point among teeth (Ledevin et al., 2010a). For each molar, the outline is described by the Cartesian coordinates of 64 points sampled at equally spaced intervals along the outline. These 64 landmarks are strongly correlated and therefore carry redundant information. To summarize this information into a lower number of variables and decrease the intensity of correlation between variables, we first performed an elliptic Fourier transform (EFT Kuhl and Giardina, 1982). The EFT provides shape variables standardized by size, the Fourier coefficients that weight the successive functions of the EFT, namely the harmonics. A study of the successive contribution of each harmonic to the description of the original outline showed that considering the first ten harmonics offered a good compromise between the number of variables and the efficient description of the outline (Ledevin et al., 2010a). Then we performed a principal component analysis of the Fourier coefficients and retained the scores on the first five principal components, which contained more than 80% of the variance (PC1=26.6%, PC2=21.6%, PC3=15.2%, PC4=7.4%, PC5=6.5%). These scores were used as phenotypic data input (the $\mathbf{y}$ data matrix) to our clustering method.

We analysed this dataset with our model first under the UFM allele frequency prior then

under the CFM prior. For each allele frequency prior, we fed the model with five types of data combination: using the georeferenced phenotypic data under the spatial model (PS), using the phenotypic data under the non-spatial model (PnS), using the georeferenced genetic data under the spatial model (GS), using the genetic data under the non-spatial model (GnS), using the georeferenced phenotypic and genetic data under the spatial model (PGS). In each case, we performed 10 independent MCMC runs of 100000 iterations discarding the first 10000 iterations as burnin.

Results

For each type of analysis, we observed an excellent congruence across the ten independent MCMC runs. The UFM and the CFM model provide qualitatively similar results with a tendency of the CFM model to return slightly larger estimates of K. While the CFM option has proven to detect finer differentiation than the UFM option (see analysis of AFLP data above), a detailed analysis and interpretation of the fine scale structure inferred by the CFM model would require extended data analysis, including some extra data still under production. We therefore focus on the results obtained under the UFM option.

In the analysis based on georeferenced phenotypic data (PS), we inferred two clusters with one cluster in the top North of Sweden (Fig. 7 top panel), all remaining samples belonging to the other cluster. These clusters correspond to the inter-specific differentiation between the red-backed vole to the North and the bank vole to the South. Analysing these data without spatial information (PnS), we also inferred also two clusters (Fig. 7 middle panel). The areas occupied by the two clusters under the PS and the PnS analyzes match in the sense that they both correspond to a top North vs. South dichotomy with a region of marked transition estimated to be along the same line in Swedish Lapland with a SW-NE orientation. In the PnS analysis, the clusters display a large amount of spatial overlap with a regular North to South cline. In

the analysis based on georeferenced genetic data (GS), we inferred the presence of four clusters. The most northern cluster corresponds to the samples identified as belonging to the top North cluster in the phenotypic clustering, and hence to the Northern red-backed vole (Fig. 7 bottom panel). The three other clusters correspond to the intra-specific structure within the bank vole. This hierarchical pattern of inter- and intra-specific differences is confirmed by estimates of inter-population differentiation provided by Fst values. The top North population attributed to the red-backed vole appears as strongly differentiated from all other populations (N Sweden vs. NE Sweden: $F_{\mathrm{ST}} = 0.15$; N vs. Central Sweden: $F_{\mathrm{ST}} = 0.19$; N Sweden vs. South Sweden: $F_{\mathrm{ST}} = 0.17$). In comparison, the differentiation is of smaller magnitude among bank vole populations (NE vs. C: $F_{\mathrm{ST}} = 0.07$; NE vs. S: $F_{\mathrm{ST}} = 0.07$; C vs. S: $F_{\mathrm{ST}} = 0.06$). Analysing these data without spatial information (GnS), we inferred four clusters whose locations match tightly those obtained under analysis GS (results not shown). In the joint analysis of georeferenced phenotypic and genotypic data (PGS), we obtained results similar to those obtained with georeferenced genetic data (results not shown).

[Figure 7 about here.]

Discussion

Summary of Approach Proposed

Main features

We have proposed the first method to date for analyzing georeferenced phenotypic and genetic data within a unified inferential framework, opening the way to combined analyses and robust comparison between markers. Our method takes as input any combination of phenotypic and genetic individual data and these data can be optionally georeferenced. Analyses can be run on phenotypic and genetic data separately or jointly. The main outputs of the method are estimates of the number of homogeneous clusters and of cluster memberships of each individual. If analyses are made on georeferenced data, the method also provides an estimate of the spatial location of each cluster which can be displayed graphically in form of continuous maps (see program documentation for details on such graphic representation).

Our approach is based on an explicit statistical model. This contrasts with model-free methods such as PAM which roughly speaking attempts to cluster individuals in order to maximize some homogeneity criterion. While such methods are fast and presumably robust to departure from specific model assumptions they are expected to behave poorly compared to methods based on an explict model that fits the data to a reasonable extent. This claim is supported by the recent study of Safner et al. (2011) in the case of spatial genetic clustering methods. In addition, because model-free methods do not rely on an explicit model, their output might be difficult to interpret or relate to biological processes.

Main results from simulation study and analysis of classic data-sets

Inference from Phenotypic Data Only: All numerical results obtained here demonstrate the good accuracy of our method and its efficiency for identifying species and/or populations boundaries. It is excellent at avoiding false positives (i.e. at reporting $\hat{K} = 1$ when K=1) and

has a clear ability to reduce the error rate when the number of variables increases. The method loses accuracy when it is given a difficult problem (i.e. when the true K is large). For a fixed number of iterations, it also has increasing difficulty to exploit fully all of the available information when the number of variables is large (cf. loss of accuracy for 50 variables compared to 20 variables), presumably due to loss of numerical efficiency in the MCMC algorithm. We also noted that MCLUST is subject to similar difficulties for large number of clusters and/or large number of variables presumably due to the existence of multiple maxima of the likelihood. In our method, this problem can be resolved to a certain extent by longer MCMC runs, an aspect not investigated in detail here. Overall, our method offers a notable improvement over the non-spatial penalized maximum likelihood method of MCLUST used under its default set of options. One factor responsible for this improvement could be that our method exploits spatial information while MCLUST does not. Results from section "Analysis of classic data of the clustering literature", where our method still provides better results than MCLUST even though the data are non-spatial, suggests this is not the sole factor. This might relate to model selection which is the second major difference between the two methods considered (Bayes vs. penalized maximum likelihood) more so bearing that MCLUST considers a broad family of covariance structure while our method assumes conditional independence.

We also stress that the numerical values characterizing the accuracy of our method have to be taken with a grain of salt since the model used to analyze the data matches exactly the model that generated them. This situation is a best case scenario and is unlikely to be strictly met in real-life cases. However, our results are informative about the potential of the method and evaluations of the iris data suggest a certain robustness of these results (see also analysis of crab morphometric data in supplementary material).

As a final note, we warn the reader unfamiliar with clustering methods against overly pes-

simistic interpretation of figure 4. From this figure, it seems that the methods lose accuracy very quickly as the the "phenotypic differentiation" decreases and are in general not so efficient. This is because detecting a hidden structure is a much harder statistical problem than testing the significance of a differentiation between two known clusters (the former involving many more parameters and hence uncertainty than the latter). More details are given in section "Power to test the significance of a known structure versus power to detect a hidden structure" of Supporting Material.

Inference from Phenotypic and Genetic Data Genetic and phenotypic data can trace different evolutionary histories, for instance phylogenetic divergence for neutral genetic markers and adaptation for a morphological structure (Renaud et al., 2007; Adams et al., 2009). Note that this is also true for any genetic marker that only traces its own evolutionary history in a phylogenetic dynamics (Turmelle et al., 2011). Confronting the structure provided by different markers emerges more and more as a way to get a comprehensive view of the dynamics and processes of differentiation among and within species. Our method, by providing a unified inferential framework for analysing different kind of data, including phenotypic ones, appears as a significant improvement for valid confrontation between data sets. Furthermore, in situations when genetic and phenotypic patterns are suspected to coincide, making inference from genetic and phenotypic data jointly has the potential to increase the power to detect boundaries between evolutionary units at different levels (populations, species).

Analysis of the *Calopogon* AFLP data-set

The ability of our model under the CFM prior to detect and classify species is excellent. This dataset has been re-analyzed by Hausdorf and Hennig (2010) who carried out a comparison of STRUCTURE, STRUCTURAMA, a method known as "field of recombination" (Doyle, 1995) and a

hybrid method mixing sequentially multidimensional scaling and model-based Gaussian clustering. The STRUCTURE program and the "field of recombination" method were not able to detect any structure. STRUCTURAMA identified only three clusters and misclassifies 44% of the samples. The hybrid method of Hausdorf and Hennig (2010) identifies 5 clusters but misclassifies 15% of the samples. Our method under the CFM prior also identifies 5 clusters but misclassifies only 5% of the samples. Under the UFM model, the results we obtain are higly consistent with those obtained with the CFM.

We also refer the reader to the Supplementary Material where we analyze AFLP data of *Veronica (pentasepalae)* from the Iberian Peninsula and Morocco produced and first analysed by Martínez-Ortega et al. (2004). The results we report there confirm the excellent performance of our method compared to the four methods investigated by Hausdorf and Hennig (2010). Finally, all the analysis carried out in the present article show that concerns of Hausdorf and Hennig (2010) against methods for dominant markers based on Hardy-Weinberg equilibrium were not grounded, provided the dominant nature of AFLP markers is taken into account at the likelihood level as we did. We suspect that the poor performances of STRUCTURE observed by Hausdorf and Hennig (2010) relate to the procedure used to estimate K (Evanno et al., 2005), as noted earlier by Waples and Gaggiotti (2006).

The *Myodes* data-set

We confronted clustering hypotheses using various data subsets with or without spatial data and with or without genetic markers or morphometric variables. This shed new lights on the population structure of *Myodes*. The pattern of phenotypic and genetic differentiation can find an interpretation in a complex pattern of contact between species and populations. The northernmost area corresponds to the narrow zone of possible overlap between *Myodes glareolus* and its close northern relative *Myodes rutilus*. Both species are difficult to recognise based on exter-

nal phenotypic characters, and impossible to identify based on common mitochondrial markers because of the introgression of *M. rutilus* mtDNA into the northern fringe of *M. glareolus* distribution. The northern cluster detected by our method corresponds most probably to the occurrence of the northern red-backed vole *M. rutilus*, that tends to differ in molar shape from its relative *M. glareolus* (Ledevin et al., 2010b).

The two analyses based on phenotypic data with and without spatial information lead to slightly different results, the former suggesting the presence of an abrupt phenotypic discontinuity in the North while the latter suggests clinal variation (Fig. 7 upper and middle panel). In absence of model fit criteria to assess the value of these two maps, we are reduced to speculate. We note however that these maps are congruent concerning the location of the main area of transition between the clusters and that the analysis based on spatial information is graphically more efficient at displaying the location of this transition. The bank vole molar shape has been shown to display a large variation even within populations, due to wear and developmental factors (Guérécheau et al., 2010; Ledevin et al., 2010b). This may render even clear cut inter-specific boundaries difficult to detect. Our georeferenced method may greatly help to make such signal emerge despite the intrinsic variability in the phenotypic markers. This suggests that our method could be viewed as an efficient generalisation of the methods aimed at detecting abrupt changes of Womble (1951) and Bocquet-Appel and Bacro (1994).

Regarding the additional clusters detected based on genetic data, the location of two of them suggests that they correspond to bank vole lineages already known in this region based on mitochondrial DNA data. Indeed, after the last ice age, Sweden has been recolonized by different populations separated several hundreds of thousand years ago coming from the South and from the North of Fennoscandia (Jaarola et al., 1999; Razzauti et al., 2009). Our new data therefore confirm the existence of two different bank vole lineages in Sweden based on mitochondrial and

now nuclear DNA markers. The existence of a fourth cluster located in Central Sweden strongly suggests that the contact zone between these two main lineages is situated in this latter region. Its origin may be attributed to hybridization between animals of the two genetic lineages. The discovery of this last cluster is new and it was never detected previously using only mitochondrial DNA marker.

Combining phenotypic and genetic data in a joint analysis (PGS) did not allow us to detect any extra structure (map not shown), possibly because beyond the inter-specific phenotypic difference corresponding to the differentiation between top North and the rest of Sweden, a cline in molar shape exists through Sweden that is roughly congruent with the genetic clusters (data not shown). It shows that the confrontation between data sets may be as informative as a joint analysis, by providing clues about the hierarchical pattern of differentiation. Morphometric clusters evidenced here inter-specific differences between red-backed and bank voles whereas based on microsatellite data, both inter- and intra-specific levels of differentiation emerged as separate clusters. The structure of genetic differentiation corroborates this interpretation. The inter-specific differentiation of the top North cluster from the rest of Sweden is indeed much stronger than the intra-specific differentiation among the bank vole populations from North-East, Central and South Sweden. Combining together both data types allows us to interpret the complex phylogeographic structure of this species and helps to distinguish differences between true species and populations within a species.

Future Extensions

Our method is based on an assumption of independence of the phenotypic variables within each cluster. This does not amount to independence between these variables globally. Indeed, the fact that phenotypic variables are sampled with cluster-specific parameters does include a correlation (similarly to the dependence structure assumed in a linear mixed model). However our

method does not deal with residual dependence not accounted for at the cluster level such as that generated by allometry. Results from simulations and classic datasets suggest that this can be partially dealt with by pre-processing the data (e.g. transforming raw data into log-shape ratio). Several other procedures may be applied for avoiding or reducing problems with covariation among phenotypic variable. For example, working on principal components rather than on raw data may help in this task. Procedures such as the Burnaby approach (Burnaby, 1966) may also allow to remove covariance structures due only to growth or other confounding factors that the user may wish to filter out. A more rigorous approach would consist in allowing the variables to covary within clusters which would also allow one to quantify these covariations.

Potential Applications

Evolutionary biology has been flooded by molecular data in the recent years. However, efficient methods to deal with phenotypic data alone are still needed when this type of data is the only available. This includes the important case of fossil data. We note that in systematic paleontology, the methods used are often simpler than those discussed in the present paper and chosen as a matter of tradition in the field rather than on objective basis. Implementing our method in a free and user-friendly program should help provide more objective methods in this context.

Our method was specifically tailored for biometric/morphometric measurements which are typically obtained at a few tens of phenotypic variables. The method proposed is therefore computer intensive and not expected to be well suited for large datasets such as expression data produced in functional genomics. However, in the situations where the scientist is able to select some variables of particular interest and reduce the dimensionality of the model (as we did for our analysis of the *Myodes* molar shape data), our method could be used and play a role in the emerging field of landscape genomics (Schwartz et al., 2010).

The sub-model for genetic data used here was presented and discussed in detail by Guillot

et al. (2005) and Guillot (2008). It has been used mostly to analyse variation and structure in neutral nuclear markers (Guillot et al., 2009) and proved useful to detect and quantify fine-scale structure typical of landscape genetics studies. The novel possibility brought here to combine it with morphometrics data might popularize this genetic model among scientists interested in larger spatial and temporal scale typical of phylogeography. In the latter field, the use of mtDNA is common. As noted earlier, the analysis of such non-recombining DNA sequence data using our method is technically possible and meaningful by recoding the various observed haplotypes as different alleles of the same locus. We stress that this approach is an expedient which incurs a considerable loss of information and that our approach should not be viewed as a substitute to those that model the genealogy of genes (including the mutational process) explicitly. Extending our model to deal with non recombining DNA in a more rigorous way is a natural direction for future work.

Our method for the combined analysis of phenotypic and genetic data can be used to assess the relative importance of random genetic drift and directional natural selection as causes of population differentiation in quantitative traits, and to assess whether the degree of divergence in neutral marker loci predicts the degree of divergence in quantitative traits (Merilä and Crnokrak, 2001). Furthermore, our method should be useful in the study of hybrid zones where, as noted by Gay et al. (2008), comparing clines of neutral genetic markers with clines of traits known to be under selection also indicates the extent to which the overall genome is under selection.

Lastly, because phenotypic and genetic markers may reflect different evolutionary or demographic history, combined analyses can help to understand the hierarchy between evolutionary units (species and populations) as shown in the *Myodes* example.

COMPUTER PROGRAM AVAILABILITY: The model presented here will be available soon as part of a new version of the R package GENELAND (version $\geq$ 4.0.0). Information will be found on the program homepage `http://www2.imm.dtu.dk/~gigu/Geneland/`.

ACKNOWLEDGMENTS: The first author is most grateful to Cino Pertoldi for discussions that prompted him to develop the model for morphometric data. Our work benefitted from discussions with Jean-Marie Cornuet and comments of Andrew J. Crawford. Part of the original data of the *Myodes* analysis belong to Bernard Lehanse's Master thesis (genetic data). We thank him for sharing these data with us. We are also grateful to Montse Martínez-Ortega and Doug Goldman for making there data available to us. This work has been supported by the French National Research Agency (project EMILE, grant ANR-09-BLAN-0145-01) and the Danish Centre for Scientific Computing (grant 2010-06-04).

References

Adams, D. C., C. M. Berns, K. H. Kozak, and J. J. Wiens. 2009. Are rates of species diversification correlated with rates of morphological evolution? Proceedings of the Royal Society of London, Biological Sciences (serie B) 276:2729–2738.

Anderson, E. 1935. The irises of the Gaspe peninsula. Bulletin of the American Iris Society 59:2–5.

Anderson, T. 1984. An introduction to multivariate statistical analysis. Probability and mathematical statistics second ed. Wiley, New York.

Balding, D. and R. Nichols. 1995. A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. Genetica 96:3–12.

Balding, D. and R. Nichols. 1997. Significant genetic correlation among Caucasians at forensic DNA loci. Heredity 78:583–589.

Banfield, J. D. and A. E. Raftery. 1993. Model-based gaussian and non-gaussian clustering. Biometrics 49:803–821.

Beadell, J. S., C. Hyseni, P. P. Abila, R. Azabo, J. C. K. Enyaru, J. O. Ouma, Y. O. Mohammed, L. M. Okedi, S. Aksoy, and A. Caccone. 2010. Phylogeography and population structure of *Glossina fuscipes fuscipes* in Uganda: Implications for control of tsetse. PLoS Neglected Tropical Diseases 4.

Biernacki, C., G. Celeux, G. Govaert, and F. Langrognet. 2006. Model-based cluster and discriminant analysis with the MIXMOD software. Computational Statistics and Data Analys 51:587–600.

608 Bocquet-Appel, J. and J. Bacro. 1994. Generalized Wombling. Systematic Biology 43:442–448.

609 Burnaby, T. P. 1966. Growth-invariant discriminant functions and generalized distances. Biomet-
610 rics 22:96–110.

611 Claude, J. 2008. Morphometrics with R. Springer.

612 Corander, J., P. Waldmann, P. Martinen, and M. Sillanpää. 2004. Baps2: Enhanced possibilities
613 for the analysis of genetic population structure. Bioinformatics 20:2363–2369.

614 Corander, J., P. Waldmann, and M. Sillanpää. 2003. Bayesian analysis of genetic differentiation
615 between populations. Genetics 163:367–374.

616 Coulon, A., G. Guillot, J. Cosson, J. Angibault, S. Aulagnier, B. Cargnelutti, M. Galan, and
617 A. Hewison. 2006. Genetics structure is influenced by lansdcape features. Empirical evidence
618 from a roe deer population. Molecular Ecology 15:1669–1679.

619 Deffontaine, V., R. Ledevin, M. C. Fontaine, J.-P. Quéré, S. Renaud, R. Libois, and J. R.Michaux.
620 2009. A relict bank vole lineage highlights the biogeographic history of the pyrenean region in
621 europe. Molecular Ecology 18:2489–2502.

622 Doyle, J. 1995. The irrelevance of allele tree topologies for species delimitation and a non-
623 topological aletrnative. Systematic Biology 20.

624 Evanno, G., S. Regnault, and J. Goudet. 2005. Detecting the number of clusters of individuals
625 using the software structure: a simulation study. Molecular Ecology 14:2611–2620.

626 Excoffier, L. and G. Heckel. 2006. Computer programs for population genetics data analysis: a
627 survival guide. Nature Review Genetics 7:745–758.

628 Falush, D., M. Stephens, and J. Pritchard. 2003. Inference of population structure using multi-
629 locus genotype data: Linked loci and correlated allele frequencies. Genetics 164:1567–1587.

630 Fisher, R. A. 1936. The use of multiple measurements in taxonomic problems. Annals of Eugenics
631 7:179–188.

632 Fontaine, M., S. Baird, S. Piry, N. Ray, K. Tolley, S. Duke, A. Birkun, M. Ferreira, T. Jauniaux,
633 A. Llavona, B. Östürk, A. Östürk, V. Ridoux, E. Rogan, M. Sequeira, U. Siebert, G. Vikingson,
634 J. Bouquegneau, and J. Michaux. 2007. Rise of oceanographic barriers in continuous popula-
635 tions of a cetacean: the genetic structure of harbour porpoises in old world waters. BMC
636 Biology 5.

637 Fraley, C. 1999. Mclust:software for model-based cluster analysis. Journal of classification 12:297–
638 306.

639 Frühwirth-Schnatter, S. 2006. Finite Mixture and Markov Switching Model. Series in Statistics
640 Springer.

641 Galarza, J., J. Carreras-Carbonell, E. Macpherson, M. Pascual, S. Roques, G. Turner, and C. Ri-
642 cod. 2009. The influence of oceanographic fronts and early-life-history traits on connectivity
643 among littoral fish species. Proceedings of the National Academy of Sciences 106:1473–1478.

644 Gay, L., P. Crochet, D. Bell, and T. Lenormand. 2008. Comparing genetic and phenotypic clines
645 in hybrid zones: a window on tension zone models. Evolution 62:2789–2806.

646 Gelman, A., J. Carlin, H. Stern, and D. Rubin. 2004. Bayesian data analysis. Chapman and Hall.

647 Gilks, W., S. Richardson, and D. Spiegelhalter, eds. 1996. Markov Chain Monte Carlo in Practice.
648 Interdisciplinary Statistics Chapman and Hall.

649 Girard, C. and S. Renaud. 2011. The species concept in a long-extinct fossil group, the conodonts.
650 Comptes Rendus Palevol 10:107–115.

651 Godsill, S. 2001. On the relationship between Markov chain Monte Carlo methods for model
652 uncertainty. Journal of Computational and Graphical Statistics 10:230–248.

653 Goldman, D. H., R. K. Jansen, C. Van Den Berg, I. J. Leitch, M. F. Fay, and M. W. Chase.
654 2004. Molecular and cytological examination of calopogon *(Orchidaceae, Epidendroideae)*:
655 Circumscription, phylogeny, polyploidy, and possible hybrid speciation. American Journal of
656 Botany 91:707–723.

657 Guérécheau, A., R. Ledevin, H. Henttonen, V. Deffontaine, J. R. Michaux, P. Chevret, , and
658 S. Renaud. 2010. Seasonal variation in molar outline of bank voles: an effect of wear? Mam-
659 malian Biology 75:311–319.

660 Guillot, G. 2008. Inference of structure in subdivided populations at low levels of genetic differ-
661 entiation. The correlated allele frequencies model revisited. Bioinformatics 24:2222–2228.

662 Guillot, G. and A. Carpentier-Skandalis. 2011. On the informativeness of dominant and co-
663 dominant genetic markers for Bayesian supervised clustering. The Open Statistics and Proba-
664 bility Journal 3:7–12.

665 Guillot, G., A. Estoup, F. Mortier, and J. Cosson. 2005. A spatial statistical model for landscape
666 genetics. Genetics 170:1261–1280.

667 Guillot, G., R. Leblois, A. Coulon, and A. Frantz. 2009. Statistical methods in spatial genetics.
668 Molecular Ecology 18:4734–4756.

669 Guillot, G. and F. Santos. 2010. Using AFLP markers and the Geneland program for the inference
670 of population genetic structure. Molecular Ecology Resources 10:1082–1084.

671 Guillot, G., F. Santos, and A. Estoup. 2008. Analysing georeferenced population genetics data
672 with Geneland: a new algorithm to deal with null alleles and a friendly graphical user interface.
673 Bioinformatics 24:1406–1407.

674 Hannelius, U., E. Salmela, T. Lappalainen, G. Guillot, C. Lindgren, U. von Döbeln, P. Lahermo,
675 and J. Kere. 2008. Population substructure in Finland and Sweden revealed by a small number
676 of unlinked autosomal SNPs. BMC Genetics 9.

677 Hausdorf, B. and C. Hennig. 2003. Biotic element analysis in biogeography. Systematic Biology
678 52:717–723.

679 Hausdorf, B. and C. Hennig. 2010. Species delimitation using dominant and codominant multi-
680 locus markers. Systematic Biology 59:491–503.

681 Hubisz, M., D. Falush, M. Stephens, and J. K. Pritchard. 2009. Inferring weak population struc-
682 ture with the assistance of sample group information. Molecular Ecology Resources 9:1322–
683 1332.

684 Jaarola, M., H. Tegelström, and K. Fredga. 1999. Colonization history in Fennoscandian rodents.
685 Biological Journal of the Linnean Society 68:113–127.

686 Joseph, L., G. Dolman, S. Donnellan, K. Saint, M. Berg, and A. Bennett. 2008. Where and when
687 does a ring start and end? testing the ring-species hypothesis in a species complex of australian
688 parrots. Proceedings of the Royal Society of London, series B 275:2431–2440.

689 Klingenberg, C. P., L. J. Leamy, E. J. Routman, and J. M. Cheverud. 2001. Genetic architecture of
690 mandible shape in mice: effects of quantitative trait loci analyzed by geometric morphometrics.
691 Genetics 157:785–802.

692 Kuhl, F. P. and C. R. Giardina. 1982. Elliptic Fourier features of a closed contour. Computer
693 Graphics and Image Processing 18:236–258.

694 Ledevin, R. 2010. La dynamique évolutive du campagnol roussâtre *(Myodes glareolus)* : structure
695 spatiale des variations morphométriques. Ph.D. thesis Université Lyon 1.

696 Ledevin, R., J. R. Michaux, V. Deffontaine, H. Henttonen, and S. Renaud. 2010a. Evolutionary
697 history of the bank vole *Myodes glareolus*: a morphometric perspective. Biological Journal of
698 the Linnean Society 100:681–694.

699 Ledevin, R., J.-P. Quéré, and S. Renaud. 2010b. Morphometrics as an insight into processes
700 beyond tooth shape variation in a bank vole population. PLoS One 5:e15470.

701 Lehanse, B. 2010. Étude génétique d'une zone de contact en Suède entre deux lignées de cam-
702 pagnols roussâtres *Myodes glareolus*. Master's thesis Université de Liège.

703 Luo, A., A. Zhang, S. Y. Ho, W. Xu, W. Shi, C. S.L., and C. Zhu. 2011. Potential efficacy of
704 mitochondrial genes for animal DNA barcoding: a case study using eutherian mamals. BMC
705 Genomics 12.

706 Marin, J., K. Mengersen, and C. Robert. 2005. Handbook of Statistics vol. 25 chap. Bayesian
707 modelling and inference on mixtures of distributions. Elsevier-Sciences.

708 Martínez-Ortega, M. M., L. Delgado, D. C. Albach, J. A. Elena-Rossello, and E. Rico. 2004.
709 Species Boundaries and Phylogeographic Patterns in Cryptic Taxa Inferred from AFLP Mark-
710 ers: *Veronica subgen. Pentasepalae (Scrophulariaceae)* in the Western Mediterranean. Sys-
711 tematic Botany 29:965–986.

712 McLachlan, G. J., D. Peel, and P. Basford, K. E. Adams. 1999. The EMMIX software for the
713 fitting of mixtures of normal and t-components. Journal of Statistical Software 4:1–4.

714 Merilä, J. and P. Crnokrak. 2001. Comparison of genetic differentiation at marker loci and quan-
715 titative traits. Journal of Evolutionary Biology 14:892–903.

716 Møller, J. and D. Stoyan. 2009. Tessellations in the Sciences: Virtues, Techniques and Applica-
717 tions of Geometric Tilings chap. Stochastic geometry and random tessellations. Springer.

718 Néraudeau. 2011. The species concept in palaeontology: Ontogeny, variability, evolution. Comptes
719 Rendus Palevol 10:71–75.

720 Nesi, N., E. Nakoumé, C. Cruaud, and A. Hassanin. 2011. DNA barcoding of African fruit bats (
721 *Mammalia, Pteropodidae*). the mitochondrial genome does not provide a reliable discrimination
722 between *Epomophorus gambianus* and *Micropteropus pusillus*. Comptes Rendus Biologies
723 334:544–554.

724 Pritchard, J., M. Stephens, and P. Donnelly. 2000. Inference of population structure using mul-
725 tilocus genotype data. Genetics 155:945–959.

726 Razzauti, M., A. Plyusnina, T. Sironen, H. Henttonen, and A. Plyusnin. 2009. Analysis of pu-
727 umala hantavirus in a bank vole population in northern Finland: evidence for co-circulation
728 of two genetic lineages and frequent reassortment between strain. Journal of General Virology
729 90:1923–1931.

730 Renaud, S., P. Chevret, and J. Michaux. 2007. Morphological vs. molecular evolution: ecology
731 and phylogeny both shape the mandible of rodents. Zoologica Scripta 36:525–535.

732 Richardson, S. and P. Green. 1997. On Bayesian analysis of mixtures with an unknown number
733 of components. Journal of the Royal Statistical Society, series B 59:731–792.

734 Robert, C. and G. Casella. 2004. Monte Carlo statistical methods. second ed. Springer-Verlag,
735 New York.

Rodríguez, F., T. Pérez, S. E. Hammer, J. Albornoz, and A. Domínguez. 2010. Integrating phylogeographic patterns of microsatellite and mtDNA divergence to infer the evolutionary history of chamois (genus *Rupicapra)*. BMC Evolutionary Biology 10:222.

Sacks, B., D. L. Bannasch, B. B. Chomel, and H. Ernst. 2008. Coyotes demonstrate how habitat specialization by individuals of a generalist species can diversify populations in a heterogeneous ecoregion. Molecular Biology and Evolution 25:1354–1395.

Safner, T., M. Miller, B. McRae, M. Fortin, and S. Manel. 2011. Comparison of Bayesian clustering and edge detection methods for inferring boundaries in landscape genetics. International Journal of Molecular Sciences 12:865–889.

Schwartz, M. K., G. Luikart, K. S. McKelvey, and S. A. Cushman. 2010. Spatial complexity, informatics, and wildlife conservation chap. Landscape genomics: A brief perspective, Pages 165–174. Springer.

Sirén, J., P. Marttinen, and J.Corander. 2011. Reconstructing population histories from single nucleotide polymorphism data. Molecular Biology and Evolution 28:673–683.

Stephens, M. 1997. Discussion of the paper by Richardson and Green "On Bayesian analysis of mixtures with an unknown number of components". Journal of the Royal Statistical Society, series B 59:768–769.

Tegelström, H. 1987. Transfer of mitochondrial DNA from the northern red-backed vole (*Clethrionomys rutilus*) to the bank vole (*C. glareolus*). Journal of Molecular Evolution 24:218–227.

Turmelle, A. S., T. H. Kunz, and M. D. Sorenson. 2011. A tale of two genomes: contrasting patterns of phylogeographic structure in a widely distributed bat. Molecular Ecology 20:357–375.

758 Waples, R. and O. Gaggiotti. 2006. What is a population? An empirical evaluation of some
759 genetic methods for indentifying the number of gene pools and their degree of connectivity.
760 Molecular Ecology 15:1419–1439.

761 Wasser, S., C. Mailand, R. Booth, B. Mutayoba, E. Kisamo, and M. Stephens. 2007. Using DNA
762 to track the origin of the largest ivory seizure since the 1989 trade ban. Proceedings of the
763 National Academy of Sciences 104:4228–4233.

764 Womble, W. 1951. Differential systematics. Science 28:315–322.

765 Yang, Z. and B. Rannala. 2010. Bayesian species delimitation using multilocus sequence data.
766 Proceedings of the National Academy of Sciences 107:9264–9269.

Appendix: Detail of MCMC Inference Algorithm

Overview

The vector of unknown parameters is $\boldsymbol{\theta} = (K, \lambda, m, \mathbf{u}, \mathbf{c}, \mathbf{f}, \tilde{\mathbf{f}}, \mathbf{d}, \boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\beta})$ which can be decomposed into $\boldsymbol{\theta}_S = (\lambda, m, \mathbf{u}, \mathbf{c})$, $\boldsymbol{\theta}_G = (\mathbf{f}, \tilde{\mathbf{f}}, \mathbf{d})$ and $\boldsymbol{\theta}_M = (\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\beta})$ blocks of parameters of the spatial, genetic and phenotypic data respectively. We alternate block updates of Metropolis-Hastings or Gibbs type and also trans-dimensional updates involving changes of K and of parts of other parameters. The updates of blocks of parameters that do not involve phenotypic data are described in Guillot et al. (2005) and Guillot (2008). We describe below updates involving phenotypic data.

Joint Updates of $(\mathbf{c}, \boldsymbol{\mu}, \boldsymbol{\sigma})$

We update jointly $\mathbf{c}, \boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ as follows. We propose a new vector $\mathbf{c}^*$ by picking two clusters at random and re-assigning some individuals of one of those two clusters to the other one at random. Then we propose $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ by sampling from the full conditional distribution $\pi(\boldsymbol{\mu}, 1/\boldsymbol{\sigma}^2|\mathbf{y}, \mathbf{c}^*)$. The Metropolis-Hastings ratio is

$$\begin{aligned} R &= \frac{\pi(\boldsymbol{\theta}^*|\mathbf{y})}{\pi(\boldsymbol{\theta}|\mathbf{y})}\frac{q(\boldsymbol{\theta}|\boldsymbol{\theta}^*)}{q(\boldsymbol{\theta}^*|\boldsymbol{\theta})} \\ &= \frac{\pi(\boldsymbol{\mu}^*, 1/\boldsymbol{\sigma}^{2^*}, \mathbf{c}^*|\mathbf{y})}{\pi(\boldsymbol{\mu}, 1/\boldsymbol{\sigma}^2, \mathbf{c}|\mathbf{y})}\frac{q(\boldsymbol{\mu}, 1/\boldsymbol{\sigma}^2|\mathbf{c})}{q(\boldsymbol{\mu}^*, 1/\boldsymbol{\sigma}^{2^*}|\mathbf{c}^*)}\frac{q(\mathbf{c}|\mathbf{c}^*)}{q(\mathbf{c}^*|\mathbf{c})} \\ &= \frac{\pi(\mathbf{c}^*|\mathbf{y})}{\pi(\mathbf{c}|\mathbf{y})}\frac{\pi(\boldsymbol{\mu}^*, 1/\boldsymbol{\sigma}^{2^*}|\mathbf{c}^*, \mathbf{y})}{\pi(\boldsymbol{\mu}, 1/\boldsymbol{\sigma}^2|\mathbf{c}, \mathbf{y})}\frac{\pi(\boldsymbol{\mu}, 1/\boldsymbol{\sigma}^2|\mathbf{c}, \mathbf{y})}{\pi(\boldsymbol{\mu}^*, 1/\boldsymbol{\sigma}^{2^*}|\mathbf{c}^*, \mathbf{y})}\frac{q(\mathbf{c}|\mathbf{c}^*)}{q(\mathbf{c}^*|\mathbf{c})} \\ &= \frac{\pi(\mathbf{c}^*|\mathbf{y})}{\pi(\mathbf{c}|\mathbf{y})}\frac{q(\mathbf{c}|\mathbf{c}^*)}{q(\mathbf{c}^*|\mathbf{c})} \end{aligned} \tag{5}$$

Interestingly, the latter expression does not depend on $(\boldsymbol{\mu}^*, \boldsymbol{\sigma}^{2^*})$, which in principle would allow us to decide whet her a new state $\boldsymbol{\theta}^*$ is accepted prior to proposing $(\boldsymbol{\mu}^*, \boldsymbol{\sigma}^{2^*})$. Unfortunately, expression (5) can not be used as $\pi(\mathbf{c}|\mathbf{y})$ is not known analytically under the present model. The ratio in equation (5) has therefore to be written as

$$R = \frac{\pi(\mathbf{y}|\boldsymbol{\mu}^*, 1/\boldsymbol{\sigma}^{2^*}, \mathbf{c}^*)}{\pi(\mathbf{y}|\boldsymbol{\mu}, 1/\boldsymbol{\sigma}^2, \mathbf{c})}\frac{\pi(\mathbf{c}^*)}{\pi(\mathbf{c})}\frac{\pi(\boldsymbol{\mu}^*, 1/\boldsymbol{\sigma}^{2^*})}{\pi(\boldsymbol{\mu}, 1/\boldsymbol{\sigma}^2)}\frac{\pi(\boldsymbol{\mu}, 1/\boldsymbol{\sigma}^2|\mathbf{y}, \mathbf{c})}{\pi(\boldsymbol{\mu}^*, 1/\boldsymbol{\sigma}^{2^*}|\mathbf{y}, \mathbf{c})}\frac{q(\mathbf{c}|\mathbf{c}^*)}{q(\mathbf{c}^*|\mathbf{c})} \tag{6}$$

which involves only analytically known expressions.

Joint Updates of $(K, \mathbf{c}, \boldsymbol{\mu}, \boldsymbol{\sigma})$

We take the same strategy as Guillot et al. (2005). The algorithm follows ideas of Richardson and Green (1997). It consists in updating K by proposing to split a cluster into two clusters or merge two clusters, in a way that complies with the spatial constraints and multivariate nature of the model. Since we use the natural prior conjugate family for parameters $\boldsymbol{\mu}^*$ and $\boldsymbol{\sigma}^*$ the full conditional $\pi(\boldsymbol{\mu}, 1/\boldsymbol{\sigma}^{2^*}|\mathbf{y}, K^*, \mathbf{c}^*)$ is available and can be used as proposal distribution as advocated for example by Godsill (2001). The acceptance ratio takes essentially the same form as in equation 6 although it is now a genuine transdimensionnal move.

Detail on Hyper-Parameters

Although we do not use exactly the same prior structure as Richardson and Green (1997), we follow largely these authors. We take $\xi_j = \sum_i y_{ij}$, $h_j = \kappa_j = 2/R_j^2$ where R_j is the range of observed values of the j-th phenotypic variable. $\beta_j|g_j, h_j, \sim \mathcal{G}(g_j, h_j)$. We also set $\alpha_j = 2$ and $g_j = 1/2$. Since $E[1/\sigma^2] = \alpha/\beta$, β represents $2/E[1/\sigma^2]$. Also $1/2h$ represents the prior mean of beta.

List of Figures

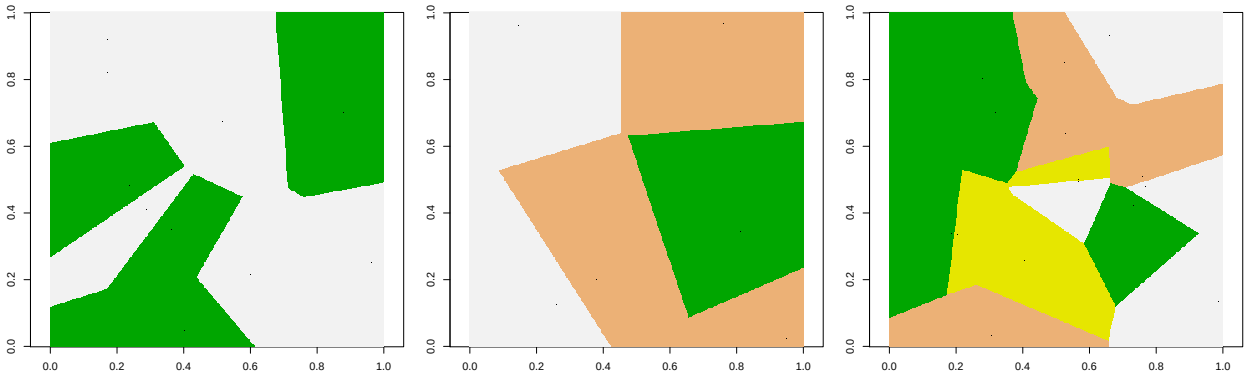


Figure 1: Examples of spatial clusters simulated from our prior model. The square represents the geographic study area. Membership of a geographical site to one of the K clusters is coded by a color. From left to right: $K = 2, 3$ and 4. A given clustering depends on K, and on the number, locations and colors (cluster memberships) of each polygon. If the prior placed on the number of polygons tends to favors low values, then each cluster tends to be made of one or only a few large areas. This is in sharp contrast with non-spatial Bayesian models which typically assume that clusterings with highly fragmented cluster areas are not unlikely.

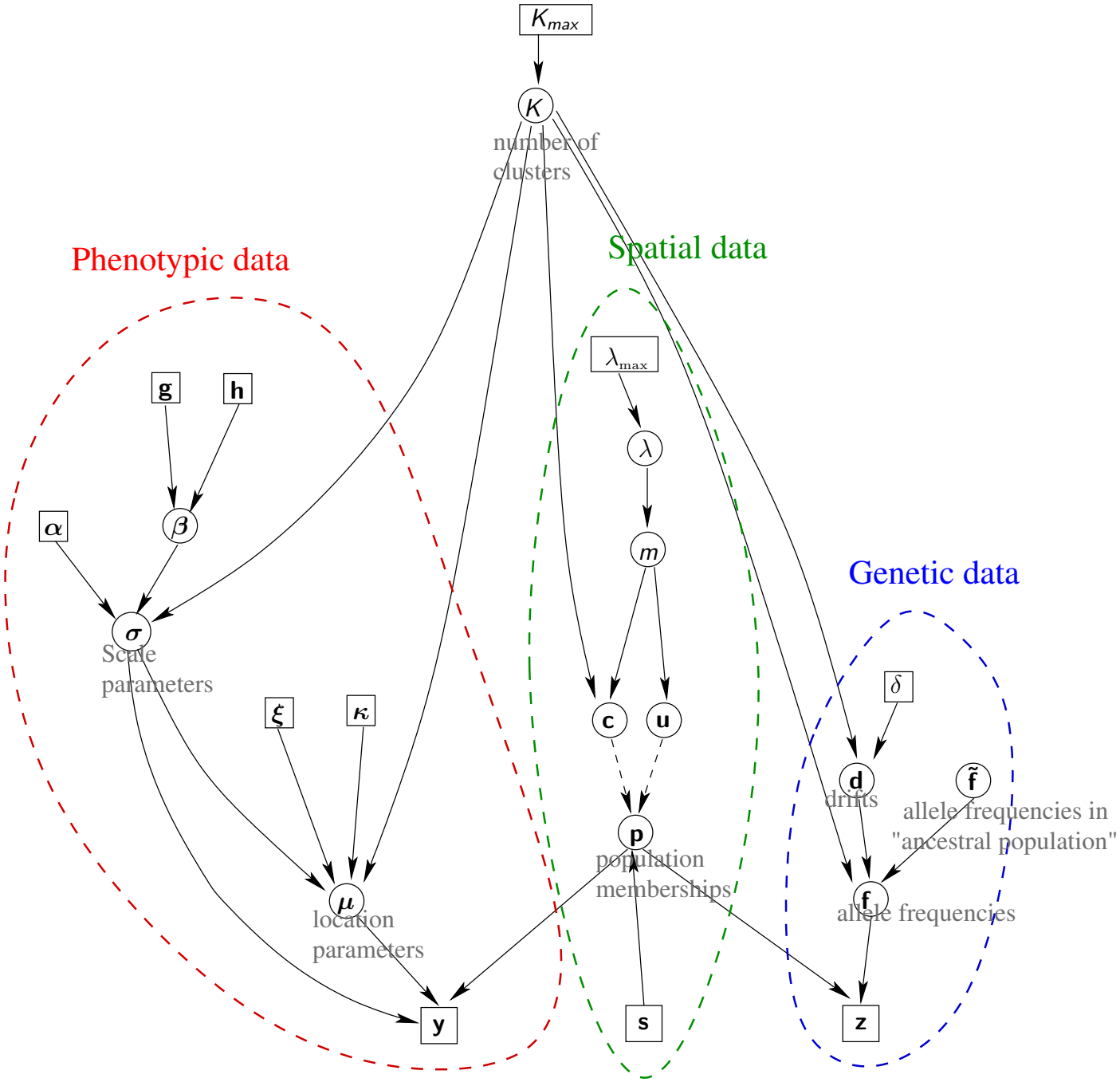
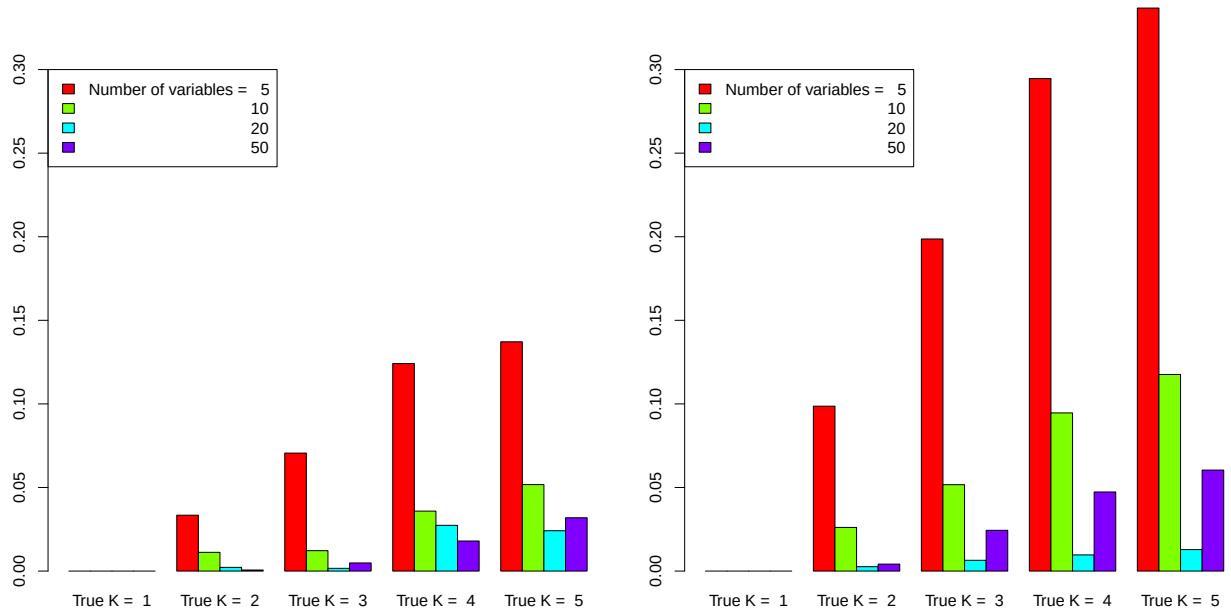


Figure 2: Graph of proposed model. Continuous black lines represent stochastic dependencies, dashed black lines represent deterministic dependencies. Boxes enclose data or fixed hyper-parameters, circles enclose inferred parameters. Bold symbols refer to vector parameters. The red, green and blue dashed lines enclose parameters relative to the phenotypic, geographic and genetic parts of the model respectively. The parameters of interest to biologists are the number of clusters K, the vector $\mathbf{p}$ which encode the cluster memberships, and possibly allele frequencies $\mathbf{f}$, mean phenotypic values $\boldsymbol{\mu}$, phenotypic variance $\boldsymbol{\sigma^2}$ which quantify the genetic and phenotypic divergence between and within clusters. Other parameters can be viewed mostly as nuisance parameters.



(a) Our method (b) Mclust

Figure 3: Classification error from simulated data. The variable plotted on the y-axis is the proportion of misclassified individuals (after correction for potential label switching issues). Each bar is obtained as an average over 500 data-sets consisting of $n = 200$ individuals. Both methods are excellent at avoiding false positives (i.e. reporting $\hat{K} = 1$ when K=1) and have a clear ability to reduce the error rate when the number of variables increases. They seem to lose accuracy in the same fashion when they are given an increasingly difficult problem (i.e. when the true K increases) and have difficulty fully exploiting all of the available information when the number of variables is large (cf. loss of accuracy for 50 variables compared to 20 variables). In the overall, under this type of simulated data, our method is typically twice as accurate as the competing method.

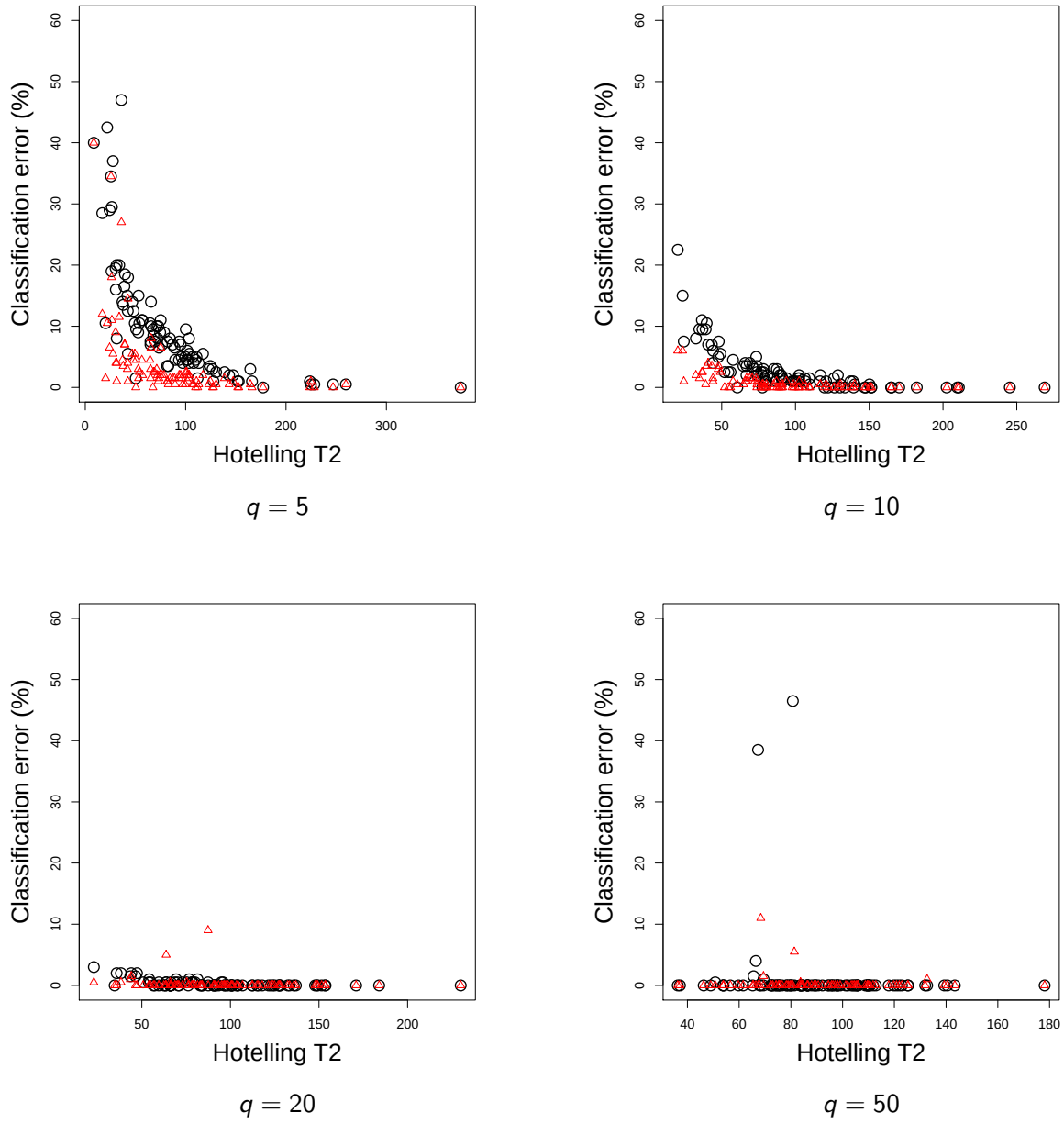
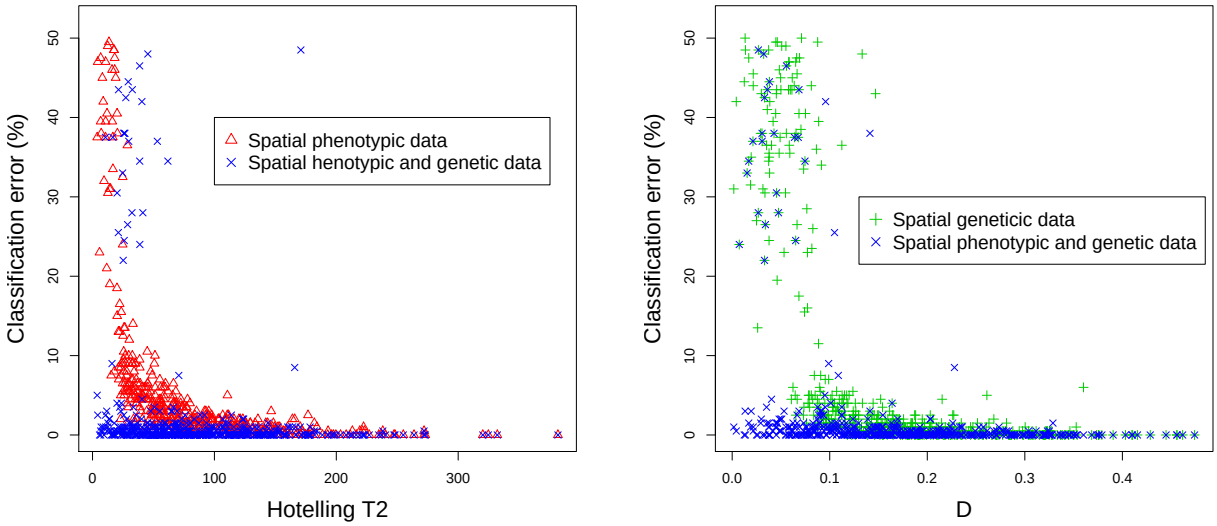


Figure 4: Classification error for simulated data-sets consisting of $K = 2$ clusters as a function of the phenotypic differentiation between the clusters. The variable plotted on the y-axis is the proportion of misclassified individuals (after correction for potential label switching issues). The variable plotted on the x-axis is the Hotelling T statistic and assesses the magnitude of the phenotypic differentiation. Our method: red triangles (△), MCLUST: black circles (○).



Average error: △ 5.2%, +8.7%, × 2.4%

Figure 5: Classification error for 500 simulated data-sets consisting of 200 individuals belonging to $K = 2$ clusters and recognized by $q = 5$ quantitative variables and $l = 10$ co-dominant loci. The variable plotted on the y-axis is the proportion of misclassified individuals using our method (after correction for potential label switching issues).

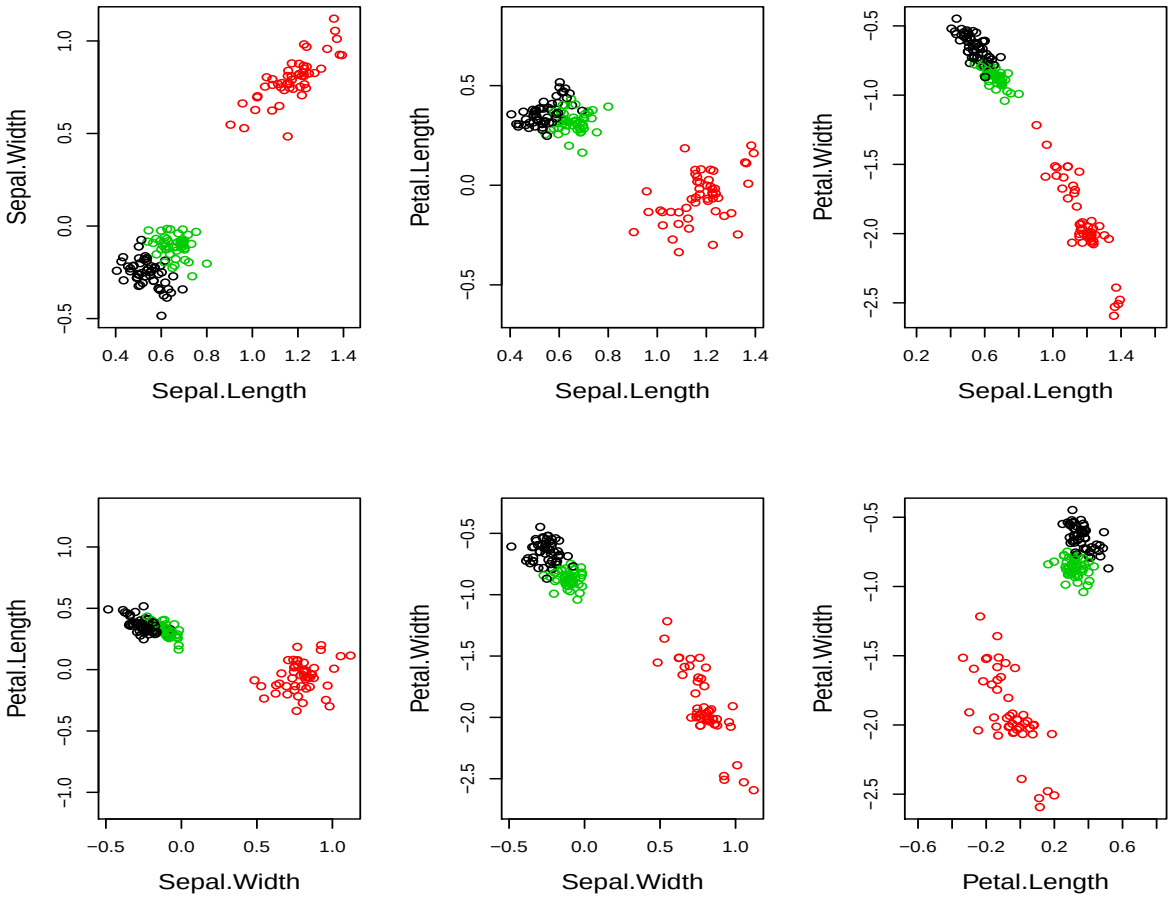


Figure 6: Pairs plots of Fisher's Iris data (transformed into log shape ratios). Colors indicate individual species estimated by our method. The true number of species (three) is correctly estimated. Only 9 out of 150 individuals are misclassified.

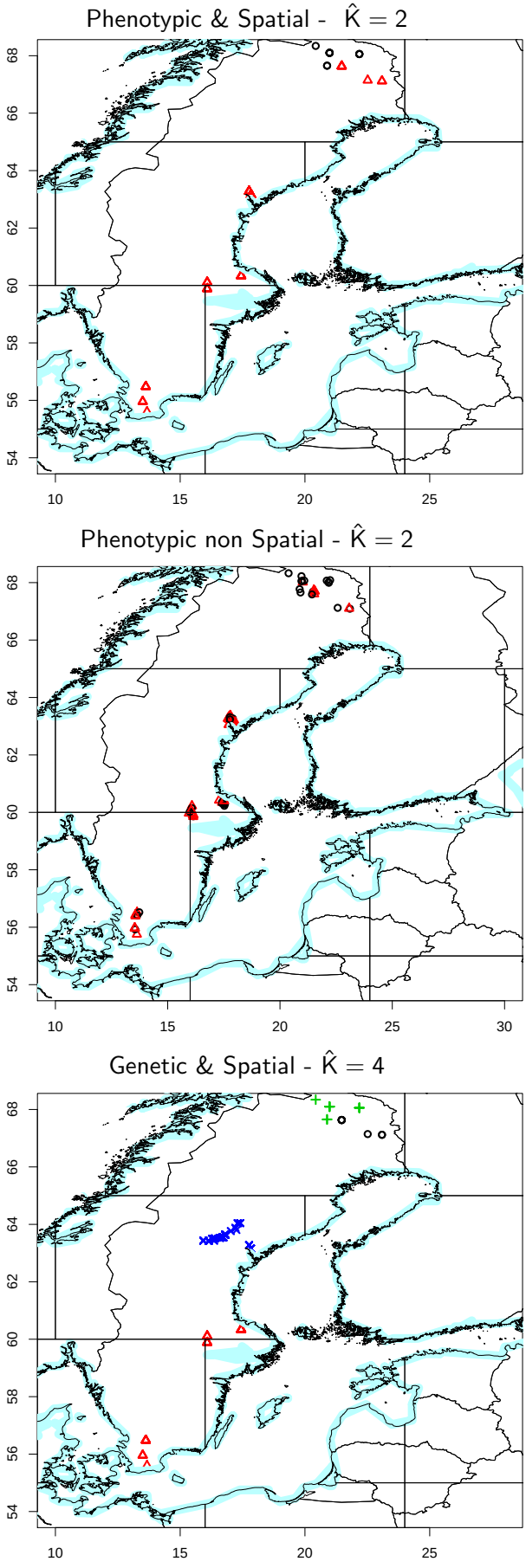


Figure 7: Population structure inferred on the bank vole data.